**GREAT LEARNING**

PGP-DSBA Online  
Feb’21 Batch

**Date**: 16/05/2021

**References**: PCA.pdf, PCA\_Multicollinearity.pdf

**Submitted By**: Prachi Gupta

**Principal component analysis(PCA) of college data**

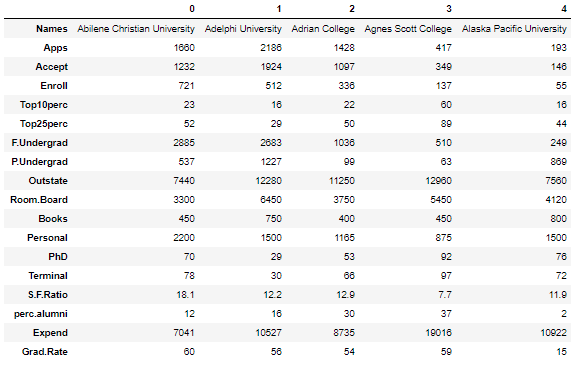
**Problem Statement:**

The dataset [Education - Post 12th Standard.csv](https://olympus.greatlearning.in/courses/36672/files/2043965/download?verifier=MGzFxmWRUU153u0YlzB1eFfPDTGhwk5VOB3xalXx&wrap=1) contains information on various colleges. You are expected to do a Principal Component Analysis for this case study according to the instructions given. The data dictionary of the 'Education - Post 12th Standard.csv' can be found in the following file: [Data Dictionary.xlsx](https://olympus.greatlearning.in/courses/36672/files/2043964/download?verifier=eZ43HPdTeZr9O9FcKB5XfuM4BxXiozT0mljarBiS&wrap=1)

1. **Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA?**

We perform EDA to get a basic information about our dataset, to serve as a base for further analysis to be performed. Let’s get a descriptive analysis of the dataset first.

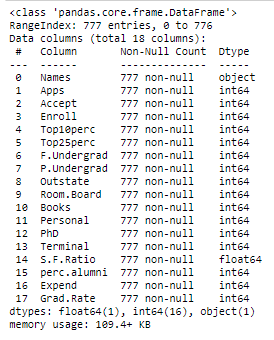
**Sample of dataset:**



**Figure: 1**

The College data set has 777 observations and 18 variables in the data set. All the variables except the ‘Names’ variable are numeric (int64 & float64 types). There are no null or duplicate values in our dataset.

**Types of variables:**



**Figure: 2**

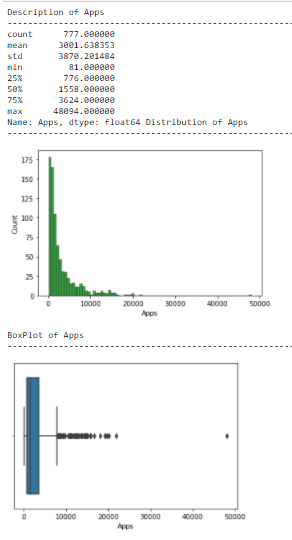
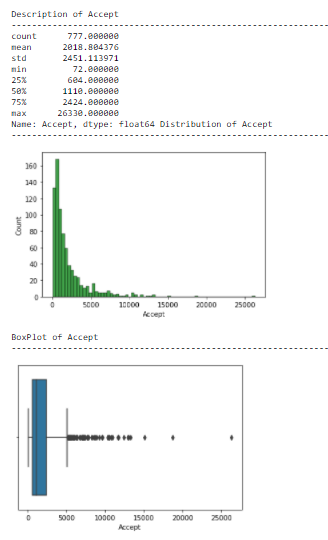
**Univariate Analysis:** Plotting Histogram, Boxplot & finding Descriptive summary for all 17 numerical columns of dataset, as all columns to be plotted are continuous in nature.

**We have eliminated 1st column:** ‘Names’ from the Univariate Analysis, as there are unique mappings for every college in data provided.

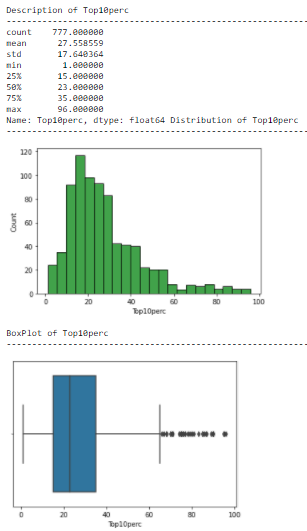
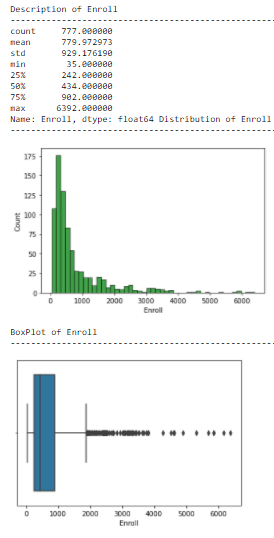
So, Count plot, Density plot wouldn’t add any value here.

**Insights:**

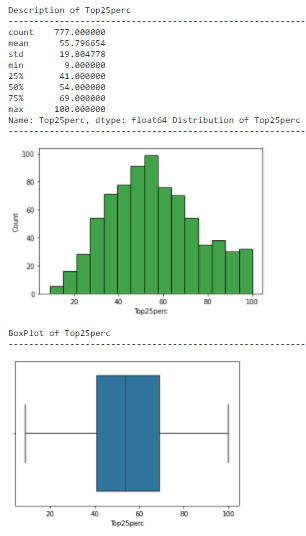
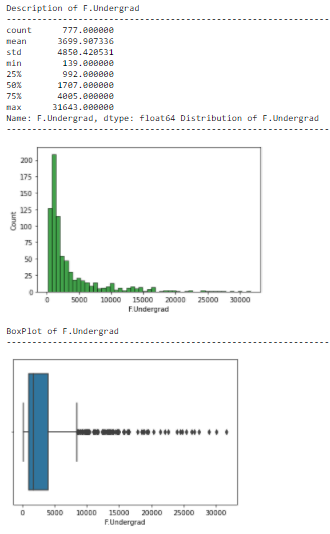
* + PhD, Terminal, Grad.Rate: **Columns having left-skewed data**
  + Expend, perc.alumni, S.F ratio, Personal, P.Undergrad, F.Undergrad, Enroll, Top10perc, Accept, Apps : **Columns having right-skewed data**
  + Room.Board, Books, Outstate, Top25perc: **Columns having data following close to normal distribution**
  + There are outliers/extreme values present for all columns except **‘**Top25perc’.
  + **‘Apps’** data has the maximum range whereas **‘S.F.Ratio’** has the minimum range of values.

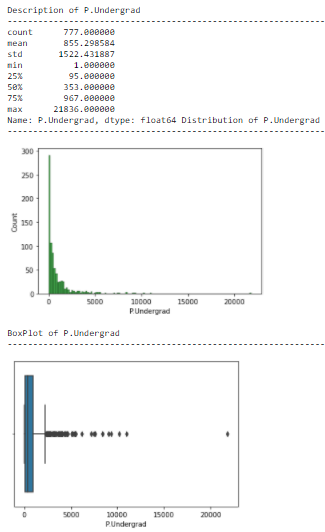
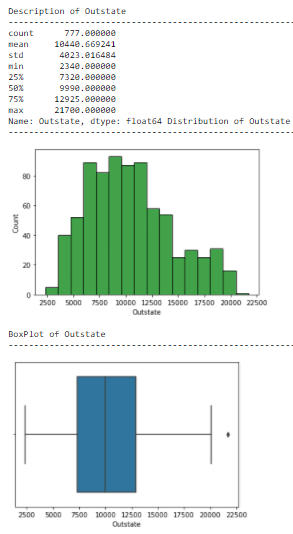
**Figure: 3 Figure: 4**



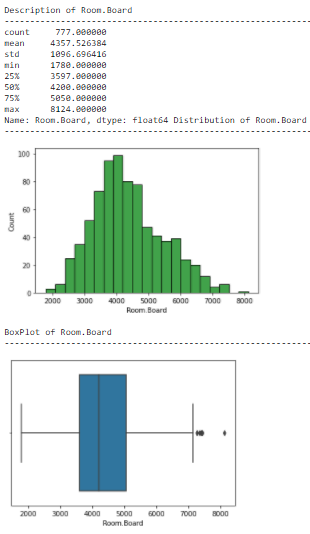
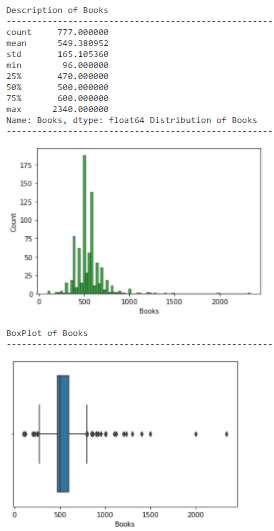
**Figure: 5 Figure: 6**

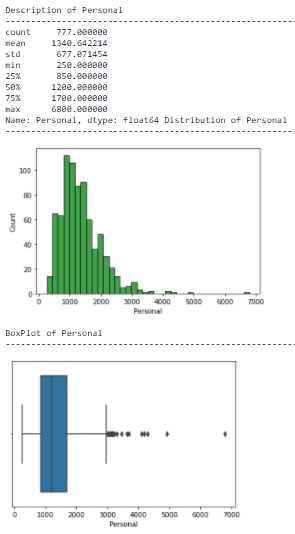
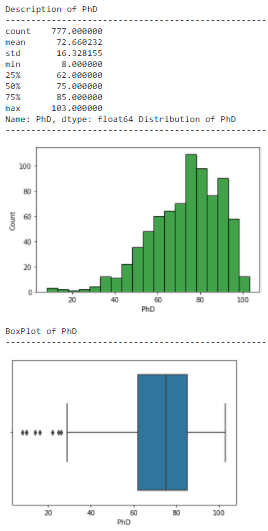
**Figure: 7 Figure: 8**

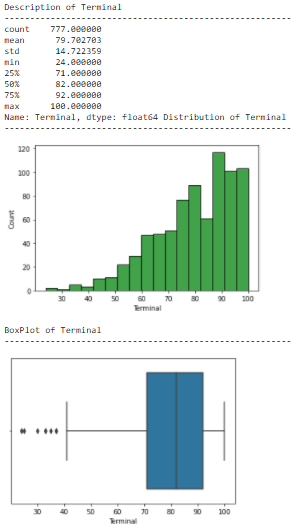
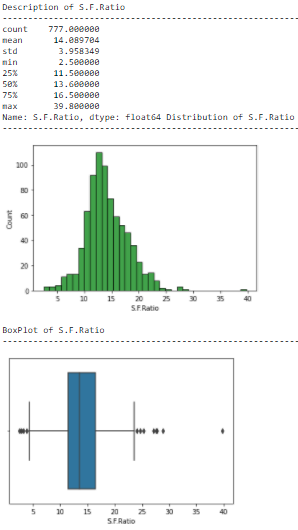
**Figure: 9 Figure: 10**

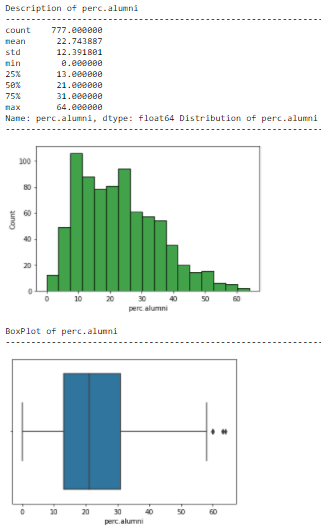
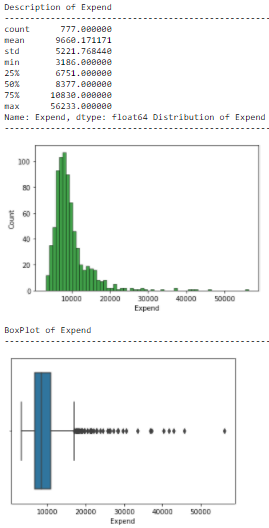
**Figure: 12 Figure: 12**

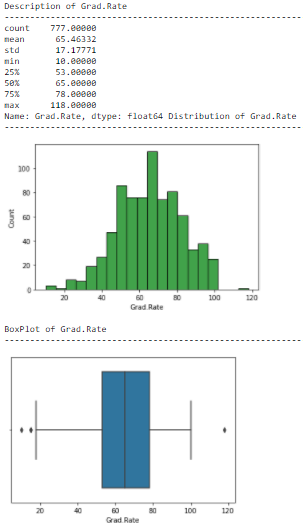
**Figure: 13 Figure: 14**

**Figure: 15 Figure: 16**

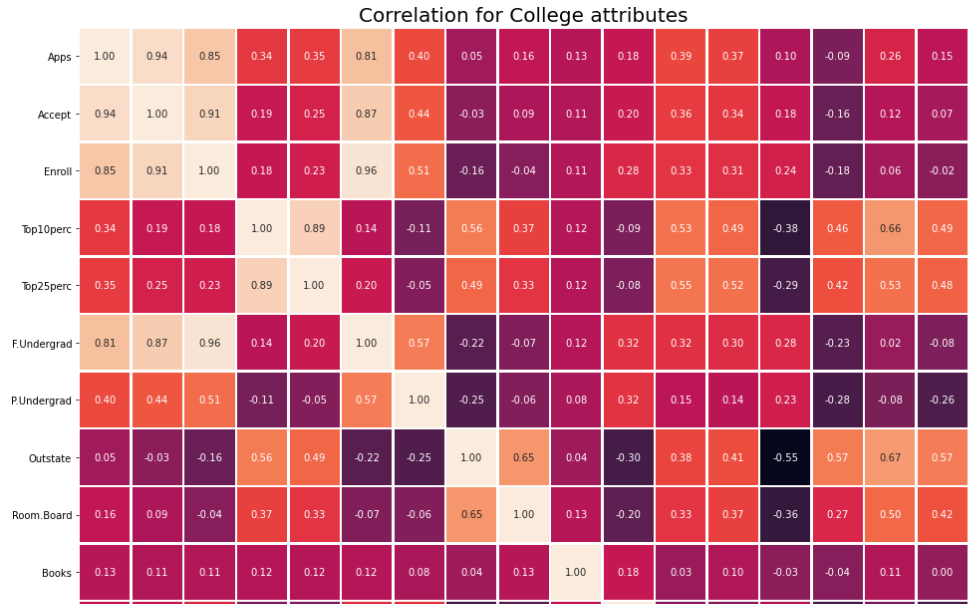
 

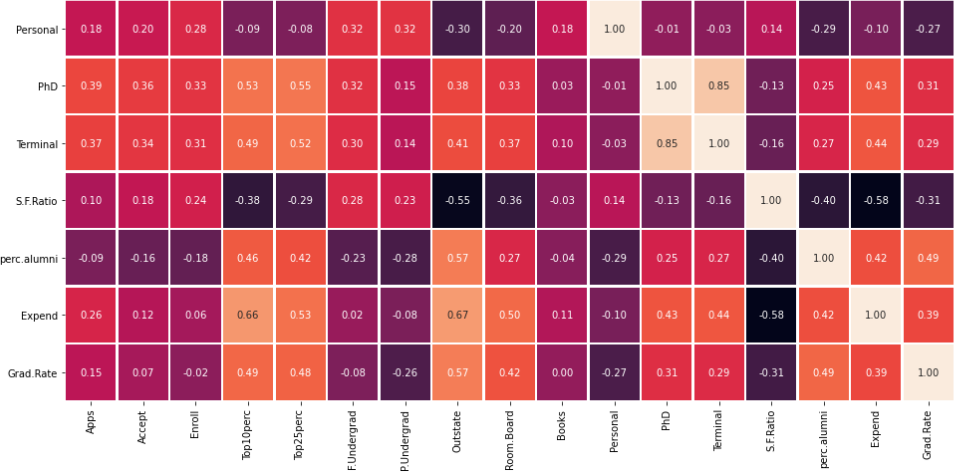
**Figure: 17 Figure: 18**

 **Figure: 19**

**Bivariate Analysis:** Plotting Pair plot (combination of scatterplots),

Heat plot is preferred for better readability of all the 17 numerical columns of dataset.





**Figure: 20**

**Insights:**

Columns ‘Enroll’ & ‘F.Undergrad’ have the strongest correlation = 0.96

Columns ‘Expend’ & ‘S.F.Ratio’ have the weakest correlation = -0.58

1. **Is scaling necessary for PCA in this case? Give justification and perform scaling.**

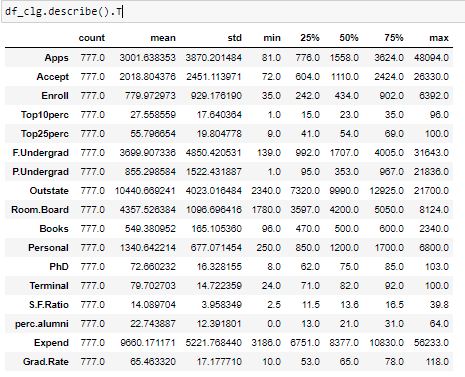
PCA tries to fetch the features with maximum variance, and the variance is high for high magnitude features.

If scaling is not done, the features with high magnitudes will weigh in a lot more than the others with low magnitudes while calculating distance.

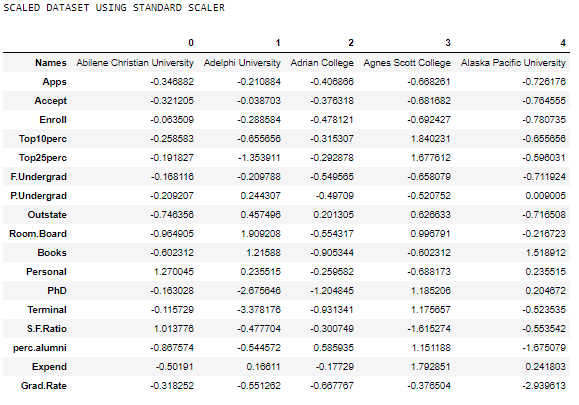
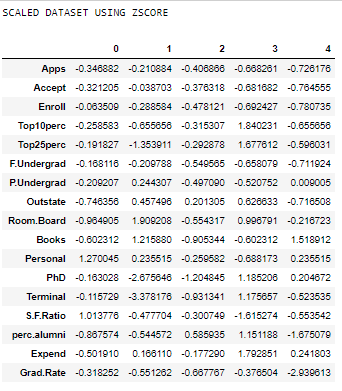
In this case, college dataset contains features highly varying in range. (For example, range of ‘Apps’ = 48013, whereas range of ‘Perc. Alumni’ = 64).

So, we must bring all features to the same level of magnitudes by performing Scaling.

The scaled data obtained ranges from **-1 to +1** for **Standard Scaler method**, whereas it ranges from **-3 to +3** for **Z-score scaling method.**



**Figure: 21**

**Figure: 22 Figure: 23**

If you look at the variables in the scaled dataset, all of them have been normalized, centered on the origins, and scaled in one scale now.

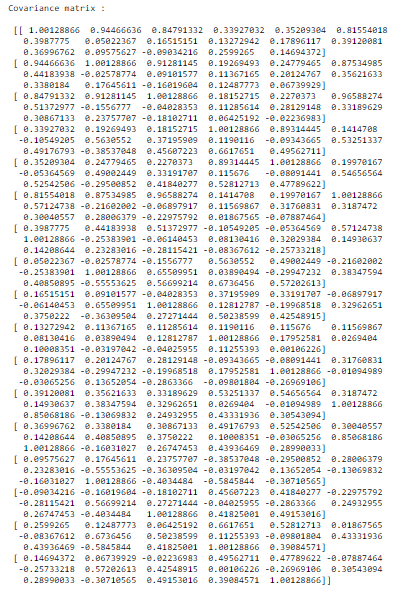
1. **Comment on the comparison between the covariance and the correlation matrices from this data [on scaled data].**

For the scaled data using z-score, standard deviation is 1. Therefore, the covariance and correlation matrices as per the below relationship must be same, ideally.

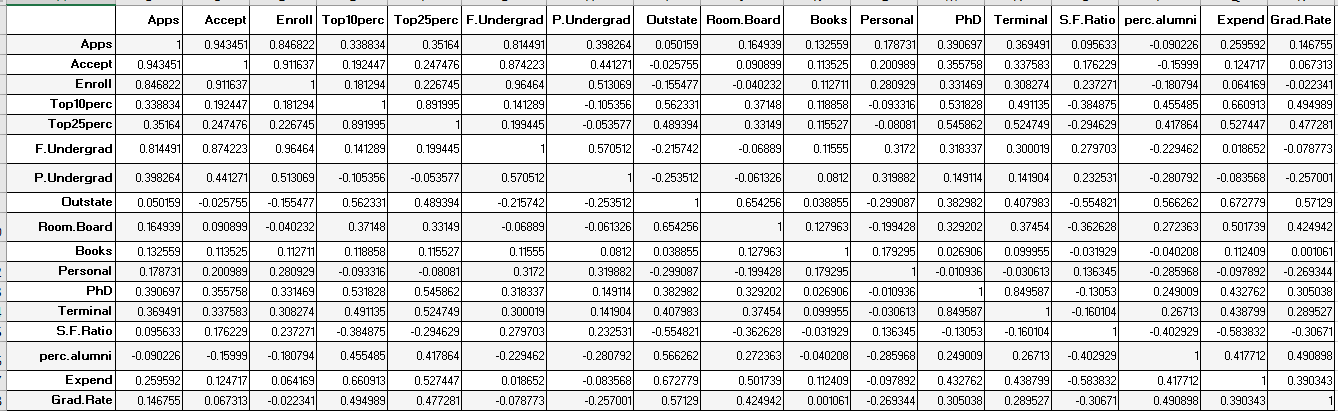
**Correlation = Covariance (x, y) / {Standard Deviation(x) \* Standard Deviation(y)}**

There can be some rounding errors while scaling & covariance calculations.

As shown in the Figure: 24 & Figure: 25 below, Covariance & Correlation matrices are approximately same, with an error in the range of 0.001.

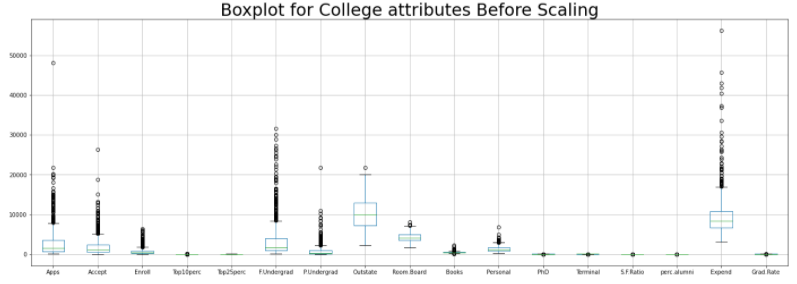


**Figure: 24** Covariance Matrix

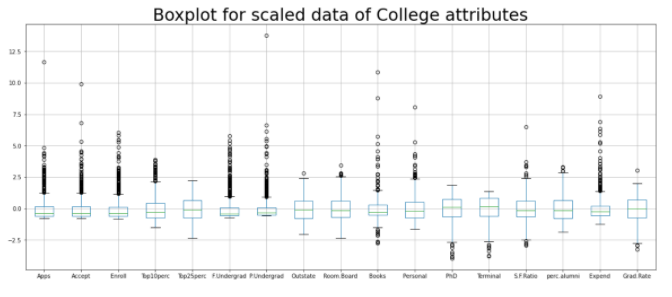
 **Figure: 25** Correlation Matrix

1. **Check the dataset for outliers before and after scaling. What insight do you derive here?**

Here are the boxplots of unscaled data **(Figure: 26, Before scaling)** v/s scaled data **(Figure: 27, After scaling).**



**Figure: 26**

 **Figure: 27**

**Insights:**

* Here data on all the dimensions are subtracted from their means to shift the data points to the origin. Hence, we get centralized data, mean=0 & Standard deviation=1, post scaling. So, the direct comparison between the data of 2 different columns becomes easier.
* As the range reduces by a considerable factor, the outliers can be visualized easily.
* In order to perform PCA on this dataset, scaled data will provide the variable with highest correlation. If not scaled, PCA will pick variable with highest variance.

1. **Extract the eigenvalues and eigenvectors.**

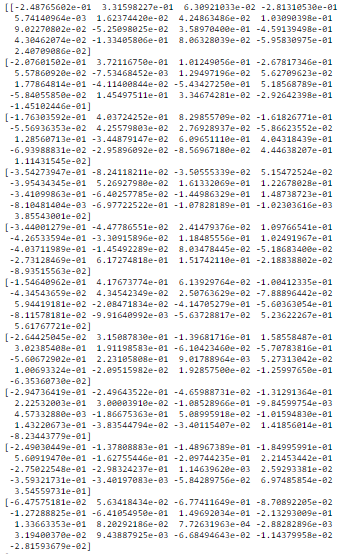
**Steps followed:**

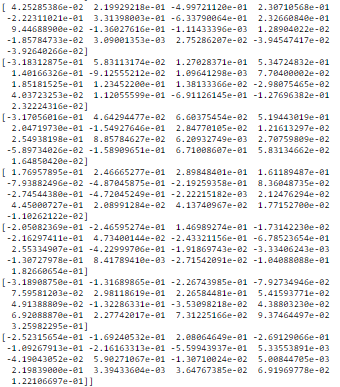
* Begin by standardizing the data. Data on all the dimensions are subtracted from their means to shift the data points to the origin. i.e. the data is centered on the origins
* Generate the covariance matrix/correlation matrix for all the dimensions
* Perform eigen-decomposition, that is, compute eigenvectors which are the principal components, and the corresponding eigenvalues which are the magnitudes of variance captured.
* Sort the eigen-pairs in descending order of eigenvalues and select the one with the largest value. This is the first principal component that covers the maximum information from the original data.

**EIGEN VALUES: Figure: 28**



**EIGEN VECTORS: Figure: 29**





1. **Perform PCA and export the data of the Principal Component (eigenvectors) into a data frame with the original features**

Following steps are followed to fetch principal components:

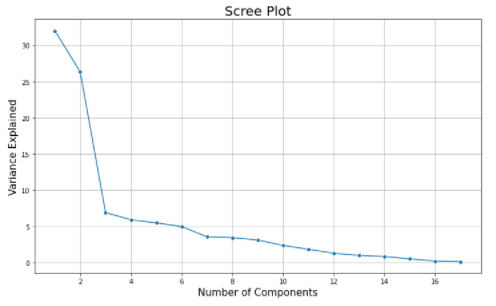
**Step 1-** Create the covariance Matrix **(Figure: 24)**

**Step 2-** Get eigen values and eigen vector (**(Figure: 28 & 29)**

Steps1 & 2 are already shown in the previous questions

**Step 3-** View Scree Plot **(Figure: 30)** to identify the number of components to be built

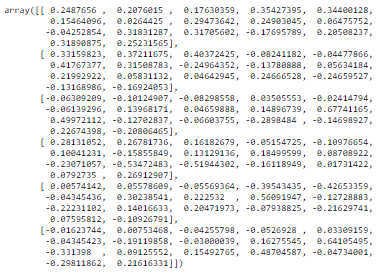
By analyzing the scree plot, we conclude that we can proceed with considering only 6 components out of 17, as a part of dimension reduction.



**Figure: 30**

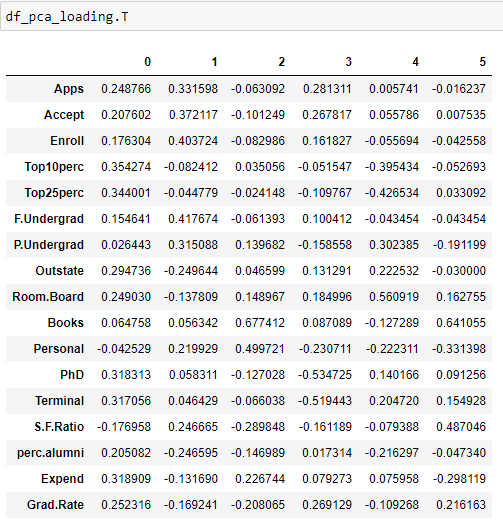
**Step 4-** Apply PCA for the number of decided components to get the loadings and component output **(Figure: 31)**

**Eigen vectors obtained post dimension reduction:**



**Figure: 31**

**Below is the exported data of the Principal Component (eigenvectors) into a data frame with the original features (Figure: 32)**



**Figure: 32**

1. **Write down the explicit form of the first PC (in terms of the eigenvectors. Use values with two places of decimals only).**

All principal components obtained are linear combination of original features.

The coefficients of linearity are determined using the Eigen vectors shown in the Figure: 31

Each row in that array corresponds to a principal component.

Thus, the first principal component, substituting the Original features is represented as:

(Considering two places of decimals, Eigen vector components are rounded off here)

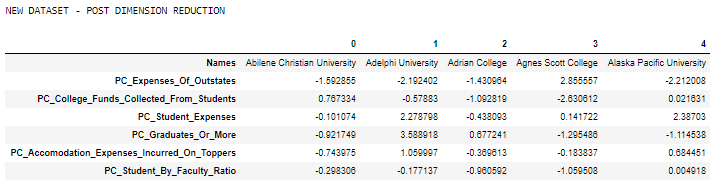
**PC0 =** 0.25\*Apps + 0.21\*Accept + 0.18\*Enroll + 0.35\* Top10perc + 0.34\* Top25perc + 0.15\* F.Undergrad + 0.03\* P.Undergrad + 0.29\*Outstate + 0.25\*Room.Board + 0.06\*Books - 0.04\*Personal + 0.32\*PhD + 0.32\*Terminal - 0.18\*S.F.Ratio + 0.21\*perc.alumni + 0.32\*Expend + 0.25\*Grad.Rate

We can fetch the values of PC0 by substituting the values (1st row of all columns of the dataset in this case, refer **Figure: 23** for scaled dataset) in the equation shown above.

Post calculations, we get the first principal component, PC0 = -1.592855

Similarly, other values can be calculated. Below data frame shows the values of PCs obtained.

This dataset can be used for any further analysis



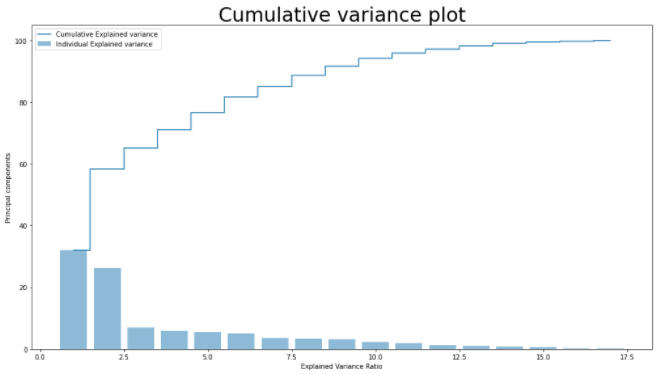
1. **Consider the cumulative values of the eigenvalues. How does it help you to decide on the optimum number of principal components? What do the eigenvector s indicate?**

Following are the cumulative values calculated for the Eigenvalues (**Figure: 33)**



**Figure: 33**

We consider only the principal components which constitute around 80% coverage of the total variance generally. As seen in **Figure: 34**, first 6 components contribute around 81% of total variance of the College dataset.

**Figure: 34**

The eigenvectors in this case indicate the direction of data variance of a particular column. Ideally, the number of Eigen vectors will be equal to the number of features present. However, there will be some directions along which the variance will be predominant.

The strength of variance along various Eigen vectors or directions is determined from Eigen values.

In this particular case, Top 6 Eigen values, when sorted in descending order, as shown in Figure: 33

Contribute to 81% of the variance.

Therefore, the Eigen vectors corresponding to these 6 Eigen values represent the Principal components.

1. **Explain the business implication of using the Principal Component Analysis for this case study. How may PCs help in the further analysis?**

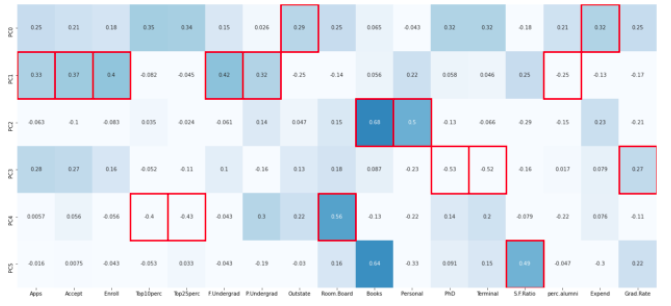
**Business Implications of using PCA:**

* The number of dimensions have been reduced significantly from 17 to 6, which contributes to approximately 81% of total variance. This helps improve the processing speed, time & performance for data wrangling algorithms.
* Some features from the dataset can have strong correlation with each other, For example., ‘Enroll’ variable has high correlation with ‘Apps’ & ‘Accept’. PCA helps reduce problem of multicollinearity, as the PCs obtained are independent of each other.
* The lower dimensional dataset thus obtained facilitates visualization & memory requirements during further analysis of data.

**The interpretation of Principal components has been done by:**

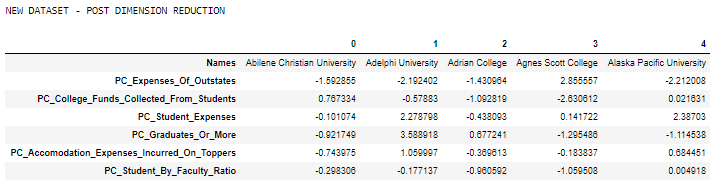
* Plotting a heatmap for the data frame having component loading against each field.
* Picking up the variables having highest correlation coefficients across each field.
* Each principal component can be renamed as per the dominant properties of the fields.
* The renamed new features are given as:
  1. PC\_Expenses\_Of\_Outstates: as the PC0(First component) has ‘Outstate’(0.29) and ‘Expend’(0.32) as dominant features. Similarly, others have been renamed
  2. PC\_College\_Funds\_Collected\_From\_Students
  3. PC\_Student\_Expenses
  4. PC\_Graduates\_Or\_More
  5. PC\_Accomodation\_Expenses\_Incurred\_On\_Toppers
  6. PC\_Student\_By\_Faculty\_Ratio

**HEATMAP SHOWING DIMENSION REDUCTION**

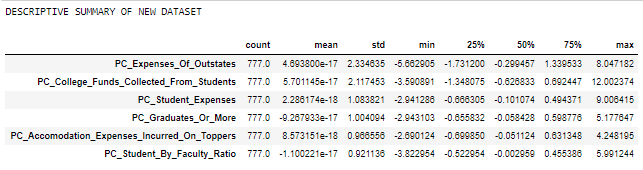


**Figure: 35**

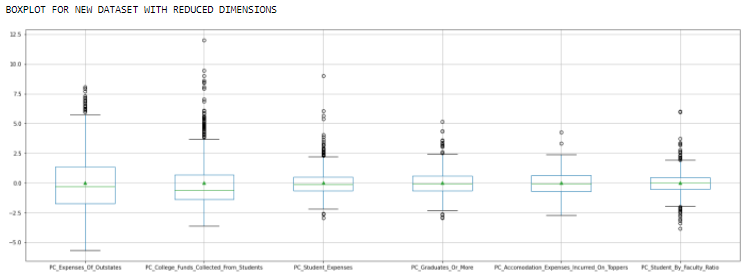
The 6 reduced variables (considering the ones having highest correlation) have been renamed, as per their business significance, and the new dataset thus formed shows in **Figure 35**.



**Figure: 36**



**Figure: 37**



**Figure: 38**

**Conclusion:**

* With help of PCA we have been able to reduce 17 numeric features into 6 components which is able to explain 81% of variance in the data.
* With help of reduced components we have been able to observe some patterns. Using some rules around business context **Figure: 38**, we observe that :

the means of columns: **‘PC\_Expenses\_Of\_Outstates’** & ‘**PC\_College\_Funds\_Collected\_From\_Students’** are relatively higher than the means of other columns.

Using the components additional rules can be derived and analyzed.

* Unsupervised learning like clustering can further be applied on the data to segment the customers based on the components created and further analyzed.